20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 1 of 36

Figure 1 - Error Recovery Architecture

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 2 of 36
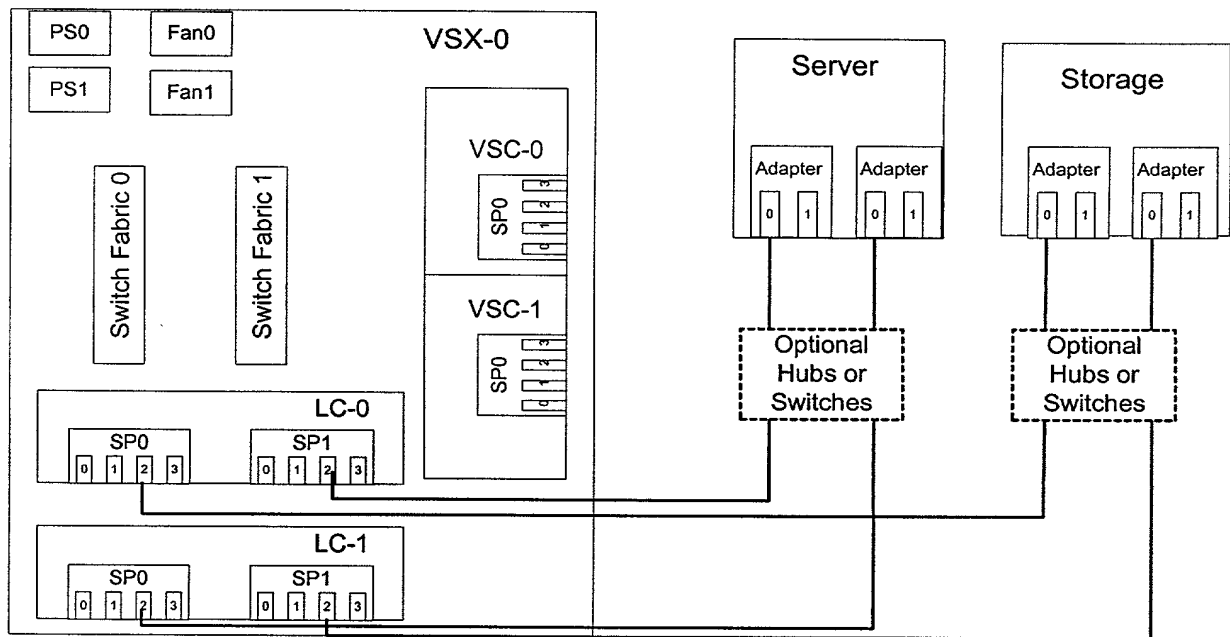
Figure 2 - Non-Fault Tolerant Configuration

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 3 of 36

Figure 3 - Fault Tolerant Configuration

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 4 of 36

Figure 4 – High Availability Configuration

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 5 of 36

Figure 5 Components of a Failover Set

Figure 6: Member Unit State Diagram

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 6 of 36
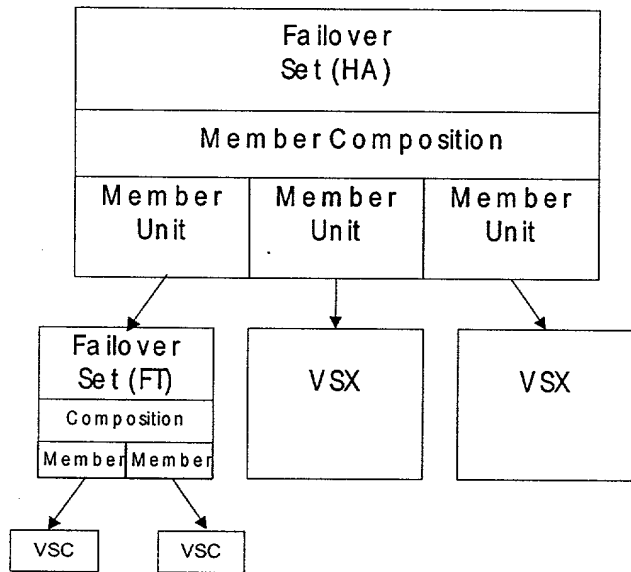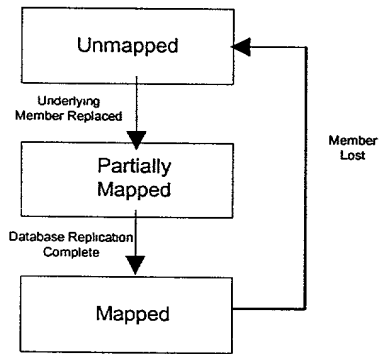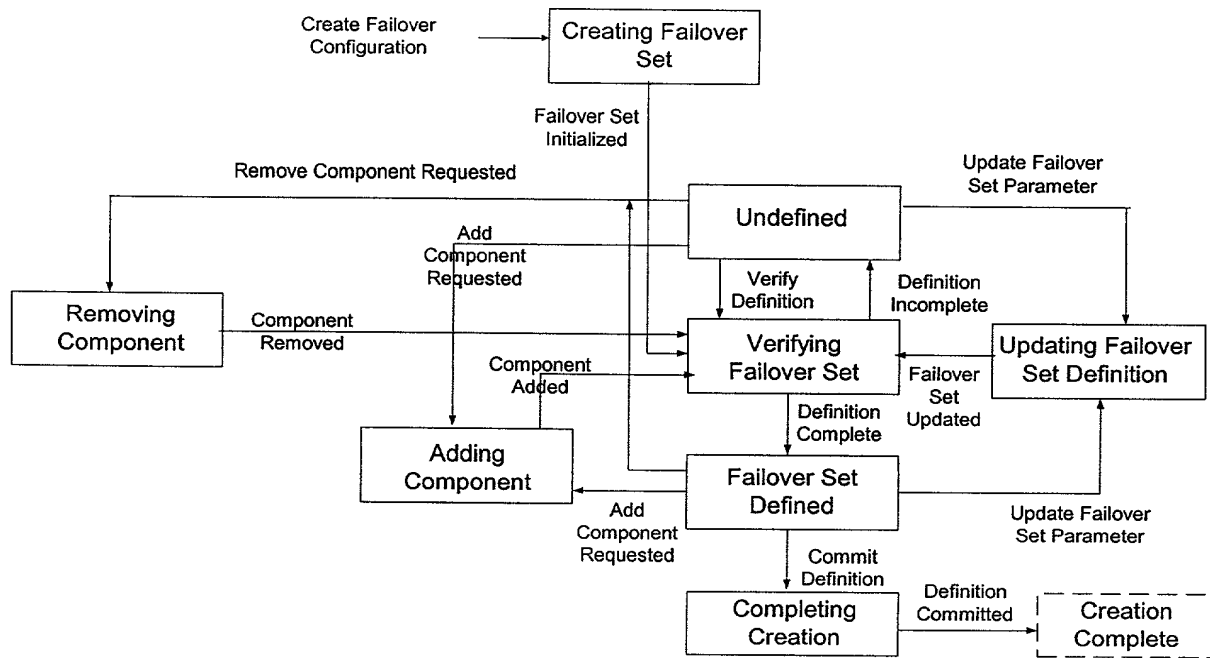
Figure 7 - Creating a Failover Set

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 7 of 36
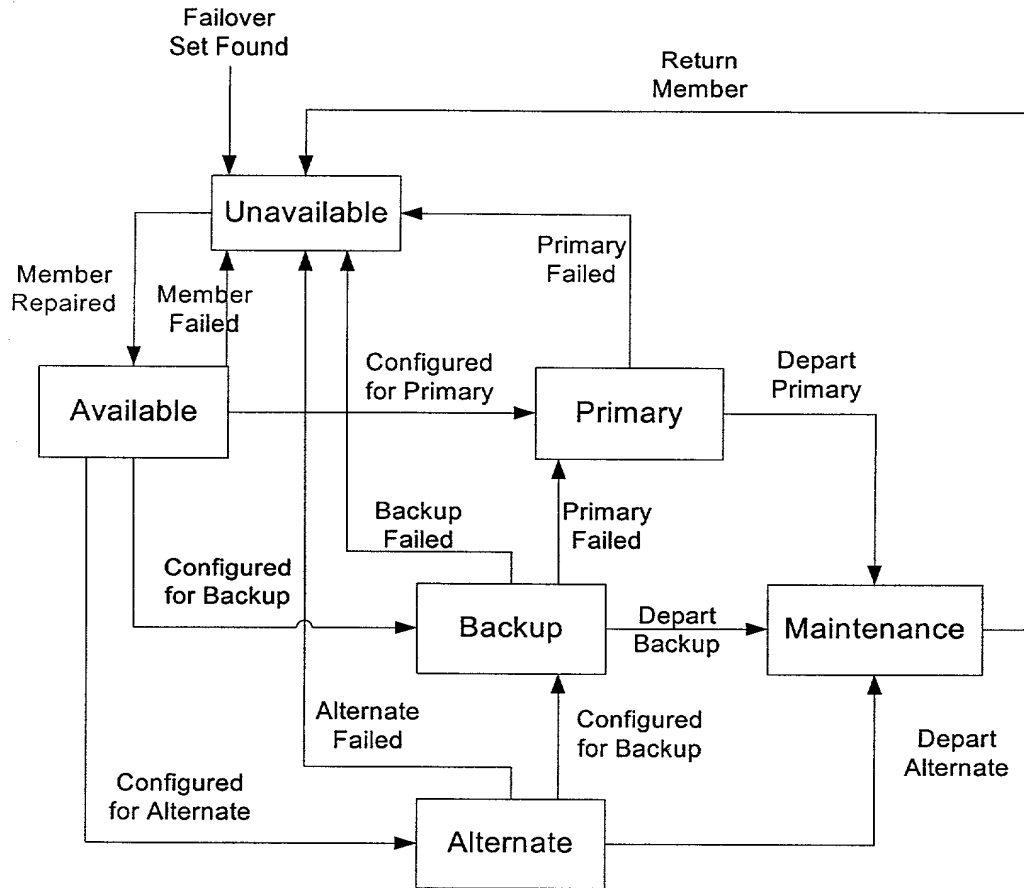
Figure 8 - Member State Diagram

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 8 of 36

Member-0	Member-1

Member-0 Repaired(COLD Boot) →

← Member-1 Repaired(COLD Boot)

← Member-1 Configured(BACKUP)

← Member-0 Repaired

Member-0 Configured(PRIMARY) →

Member-0 COLD Boot and Member-1 COLD Boot means Assume default DB role (e.g. Primary)

Member-0 COLD Boot and Member-1 COLD Boot means Assume default DB role (e.g. Backup)

Figure 9 – Member Arbitration for COLD Boot

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 9 of 36

Member-0 | Member-1

Member-0 Repaired(WARM Boot)

Member-1 Repaired(WARM Boot)

Member-1 Configured(PRIMARY)

Member-0 Repaired

Member-0 Configured(BACKUP)

Member-0 WARM Boot and Member-1 WARM Boot means Assume previous DB role (e.g. Backup)

Member-0 WARM Boot and Member-1 WARM Boot means Assume previous DB role (e.g. Primary)

Figure 10 – Member Arbitration for WARM Boot

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 10 of 36

Member-0	Member-1

Member-0 Repaired(COLD Boot)

Member-1 Repaired(WARM Boot)

Member-1 Configured(PRIMARY)

Member-0 Repaired

Member-0 Configured(BACKUP)

Member-0 COLD Boot and Member-1 WARM Boot means Assume remaining DB role (e.g. Backup)

Member-0 COLD Boot and Member-1 WARM Boot means Assume previous DB role (e.g. Primary)

Figure 11 - Member Arbitration for Mixed Boot

Old State	Event					
	Mi Repaired	Mj Repaired	Mi Configured	Mj Configured	Mi Failed	Mj Failed
1. {Mi,Mj} Unavail, {} Avail, {} Primary, {} Backup*	New St: 3 Action: A	New St: 2 Action: B			New St: 1 Action: S	New St: 1 Action: T
2. {Mi} Unavail, {Mj} Avail, {} Primary, {} Backup	New St: 4 Action: C			New St: 8 Action: D	New St: 2 Action: S	
3. {Mj} Unavail, {Mi} Avail, {} Primary, {} Backup		New St: 4 Action: E	New St: 9 Action: F			New St: 3 Action: T
4. {} Unavail, {Mi, Mj} Avail, {} Primary, {} Backup			New State: 7 Action: G	New St: 6 Action: H		
5a. {} Unavail, {} Avail, {Mi} Pri, {Mj} Backup					New St: 8 Action: I	New St: 9 Action: J
5b. {} Unavail, {} Avail, {Mj} Pri, {Mi} Backup					New St: 8 Action: I	New St: 9 Action: J
6. {} Unavail, {Mi} Avail, {Mj} Pri, {} Backup	New St: 6 Action: K		New St: 5a, 5b Action: L			New St: 3 Action: M
7. {} Unavail, {Mj} Avail, {Mi} Pri, {} Backup		New St: 7 Action: N		New St: 5a, 5b Action: O	New St: 2 Action: P	
8. {Mi} Unavail, {} Avail, {Mj} Pri, {} Backup	New St: 6 Action: C					New St: 1 Action: Q
9. {Mj} Unavail, {} Avail, {Mi} Pri, {} Backup		New St: 7 Action: E			New St: 1 Action: R	
* Initial State						

Figure 12 – 2 Member State Table

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 12 of 36

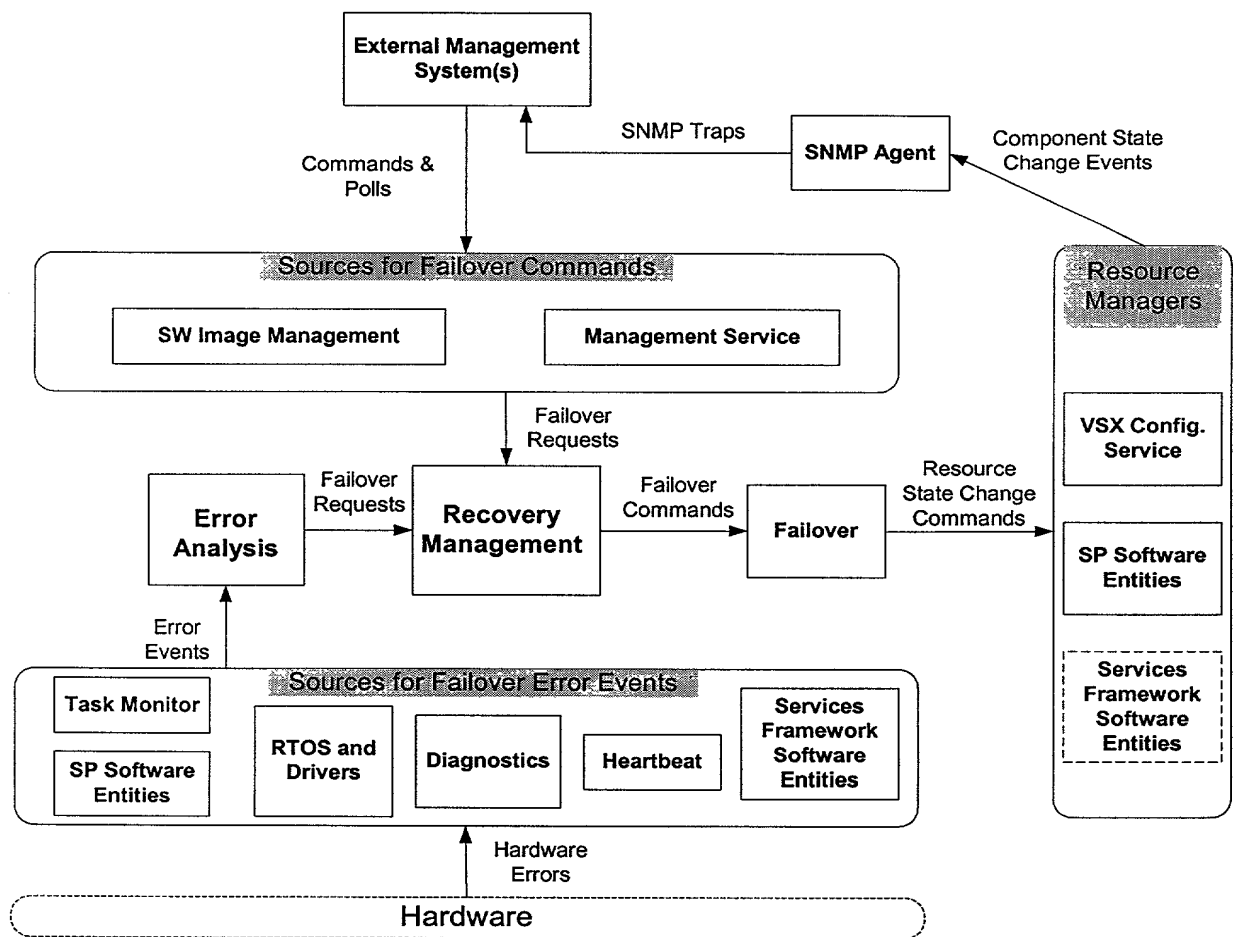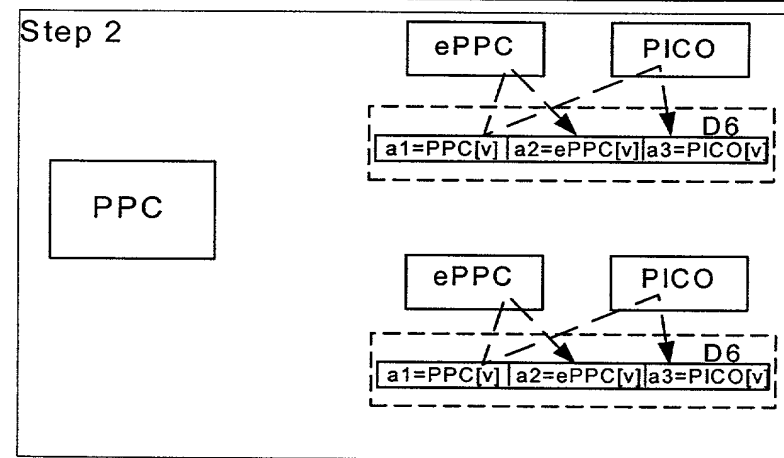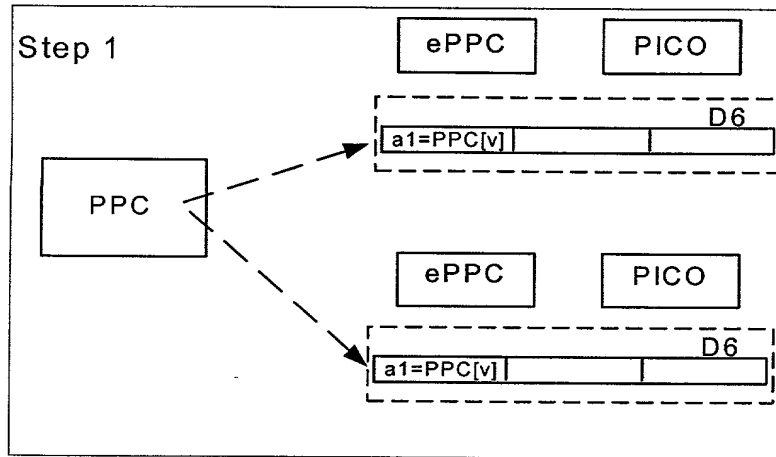| Action Routines	Description
1	1. Send "Mi repaired" to Mj, if Mj is not failed. 2. Set timer to send "Mi repaired" to Mi
2	1. Send "Mj repaired" to Mi, if Mi is not failed. 2. Set timer to send "Mj repaired" to Mj
A	1. If Mi and configured send "Mi configured" to Mj. 2. Set timer to send "Mi configured" to Mi. 3.
B	1. If Mj and configured send "Mj configured" to Mi. 2. Set timer to send "Mj configured" to Mj. 3.
C	1. If Mj, echo event back to Mi. 2. If Mi and configured send "Mi configured" to Mj. 3. Set timer to
D	1. If Mj become Primary. 2. Otherwise, nop.
E	1. If Mi, echo event back to Mj. 2. If Mj and configured send "Mj configured" to Mi. 3. Set timer to
F	1. If Mi become Primary. 2. Otherwise, nop.
G	1. If Mi become Primary. 2. Otherwise, echo event back to Mi.
H	1. If Mj become Primary. 2. Otherwise, echo event back to Mj.
I	1. If Mj become Primary. 2. If Mi become Backup.
J	1. If Mi become Primary. 2. If Mj become Backup.
K	1. If Mj echo event back to Mi. 2. Otherwise, nop
L	1. If Mj determine Member Role. 2. Send "Mi configured" to Mi when done. 3. If Mi determine
M	1. If Mj perform Fail-Stop processing. 2. Send "Mj Failed" to Mi. 3. Otherwise become Primary after
N	1. If Mi echo event back to Mj. 2. Otherwise, nop
O	1. If Mi determine Member role. 2. Send "Mj configured" to Mj when done. 3. If Mj determine
P	1. If Mi perform Fail-Stop processing. 2. Send "Mi Failed" to Mj. 3. Otherwise become Primary after
Q	1. If Mj perform Fail-Stop processing for Mj 2. Otherwise nop.
R	1. If Mi perform Fail-Stop processing for Mi. 2. Otherwise nop.
S	1. Perform Fail-Stop processing for Mi
T	1. Perform Fail-Stop processing for Mj

Figure 13 –Action Routines for a 2 Node Configuration

Figure 14: Served Failover Set State Machine Diagram
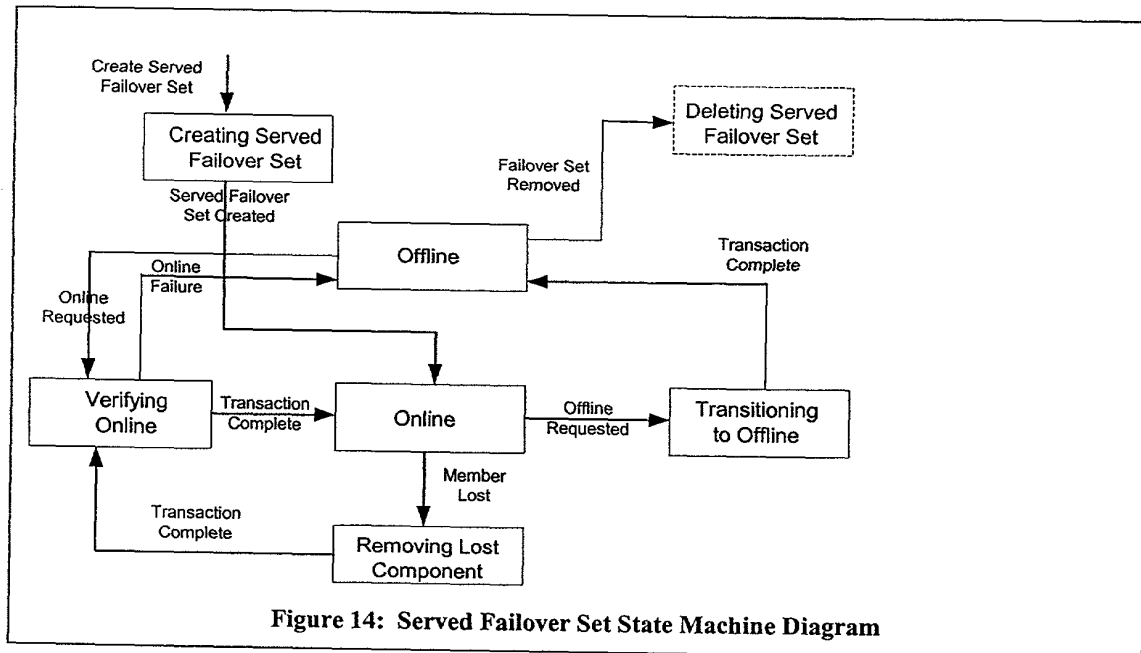
Figure 15 - Fault Detection and Analysis Architecture

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 15 of 36

Step 1

ePPC   PICO

D6
a1=PPC[v]

PPC

ePPC   PICO

D6
a1=PPC[v]

Step 2

ePPC   PICO

D6
a1=PPC[v] | a2=ePPC[v] | a3=PICO[v]

PPC

ePPC   PICO

D6
a1=PPC[v] | a2=ePPC[v] | a3=PICO[v]

Step 3

majority(a1,a2,a3) = majority(v,v,v) = v, No faults

Figure 16 - No faults

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 16 of 36

Step 1

ePPC   PICO

D6

a1=PPC[x]

PPC

ePPC   PICO

D6

a1=PPC[y]

Step 2

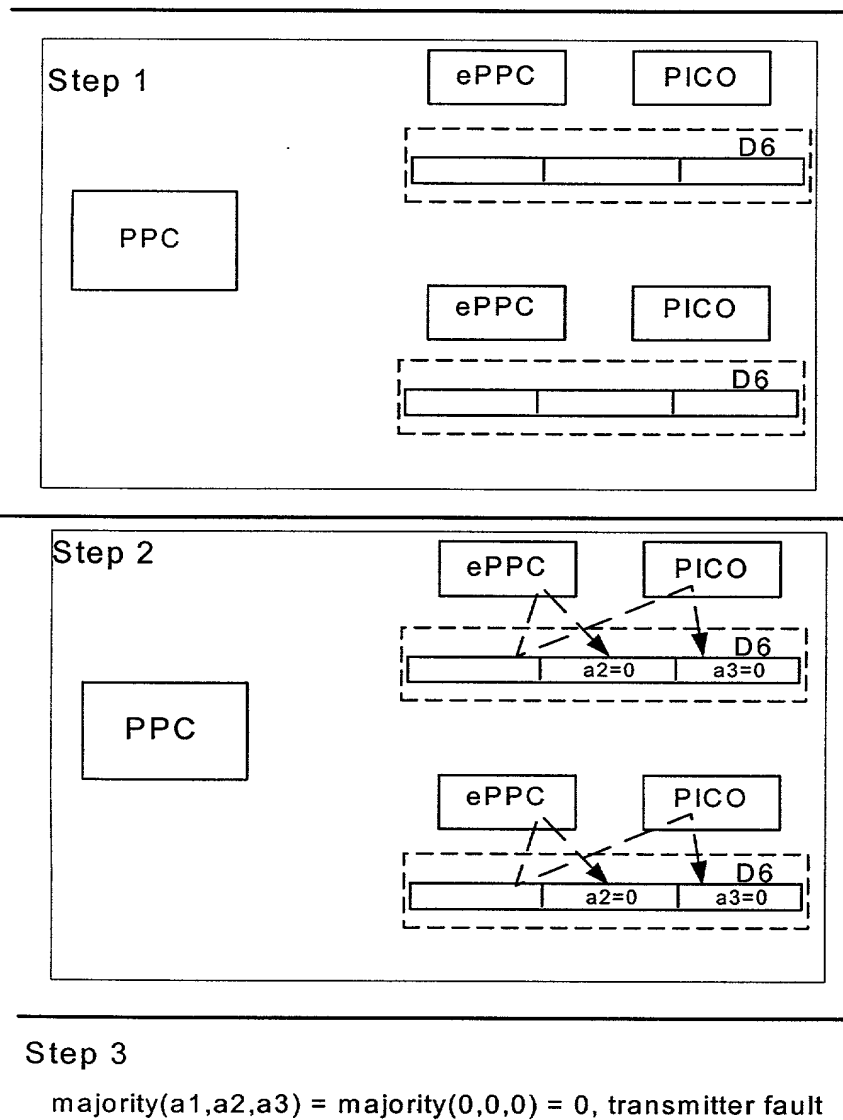ePPC   PICO

D6

a1=PPC[x]   a2=0   a3=0

PPC

ePPC   PICO

D6

a1=PPC[y]   a2=0   a3=0
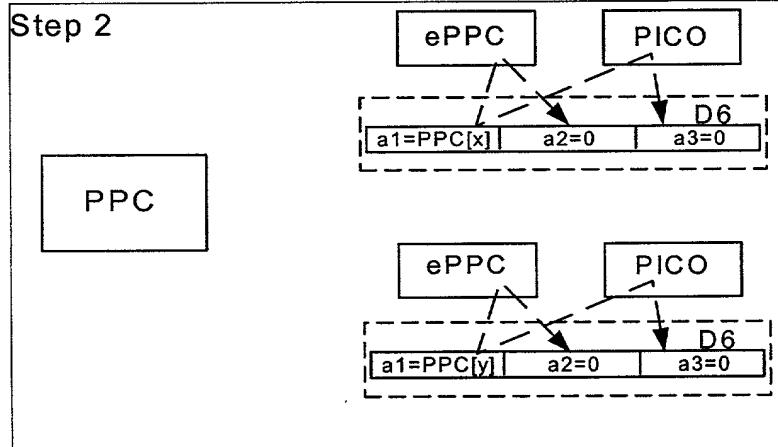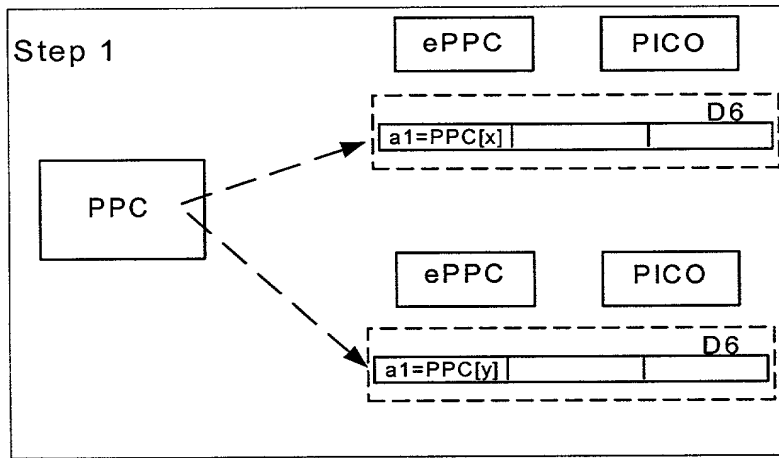
Step 3

majority(a1,a2,a3) = majority(x,0,0) = 0, transmitter fault

Figure 17 - Transmitter fault (sends a bad value)

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 17 of 36

Step 1

ePPC   PICO

D6

PPC

ePPC   PICO

D6

Step 2

ePPC   PICO

a2=0   a3=0   D6

PPC

ePPC   PICO

a2=0   a3=0   D6

Step 3

majority(a1,a2,a3) = majority(0,0,0) = 0, transmitter fault

Figure 18 - Transmitter fault (doesn't send a value)

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 18 of 36

Step 1

ePPC   PICO

D6

a1=PPC[v]

PPC

ePPC   PICO

D6

a1=PPC[v]

Step 2

ePPC   PICO

D6

a1=PPC[v] | a2=ePPC[y] | a3=PICO[v]

PPC

ePPC   PICO

D6

a1=PPC[v] | a2=ePPC[v] | a3=PICO[v]
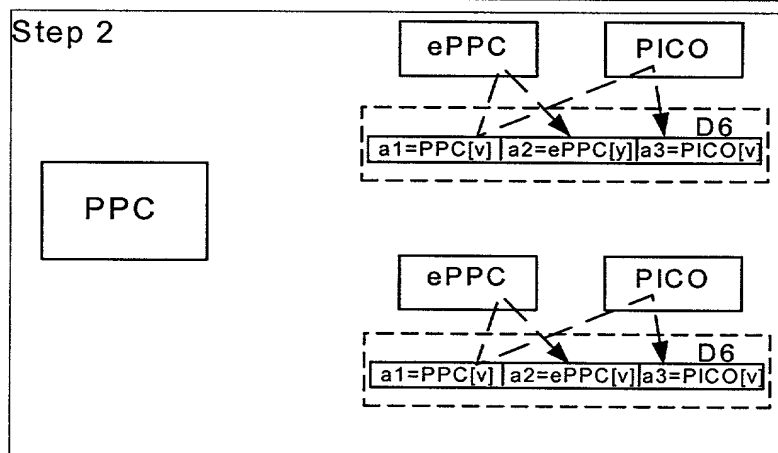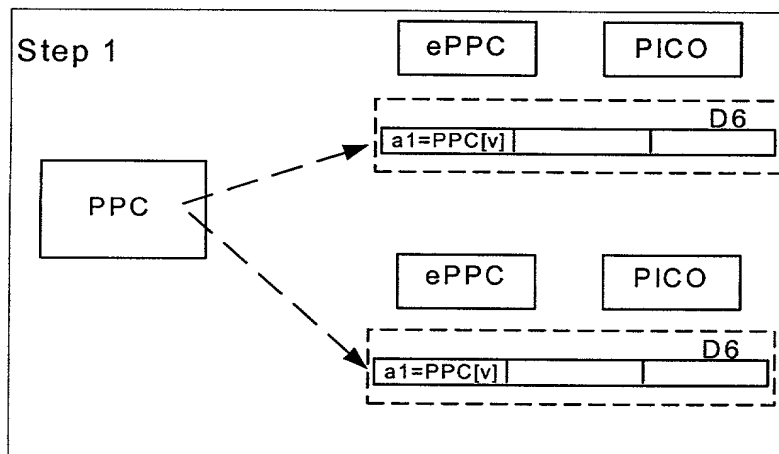
Step 3

majority(a1,a2,a3) = majority(v,y,v) = v, Receiver fault

Figure 19 – Receiver fault (relays wrong value)

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors. Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 19 of 36

Step 1

ePPC   PICO

D6
a1=PPC[v]

PPC

ePPC   PICO

D6
a1=PPC[v]

Step 2

ePPC   PICO

D6
a1=PPC[v]   0   a3=PICO[v]

PPC

ePPC   PICO

D6
a1=PPC[v]   0   a3=PICO[v]
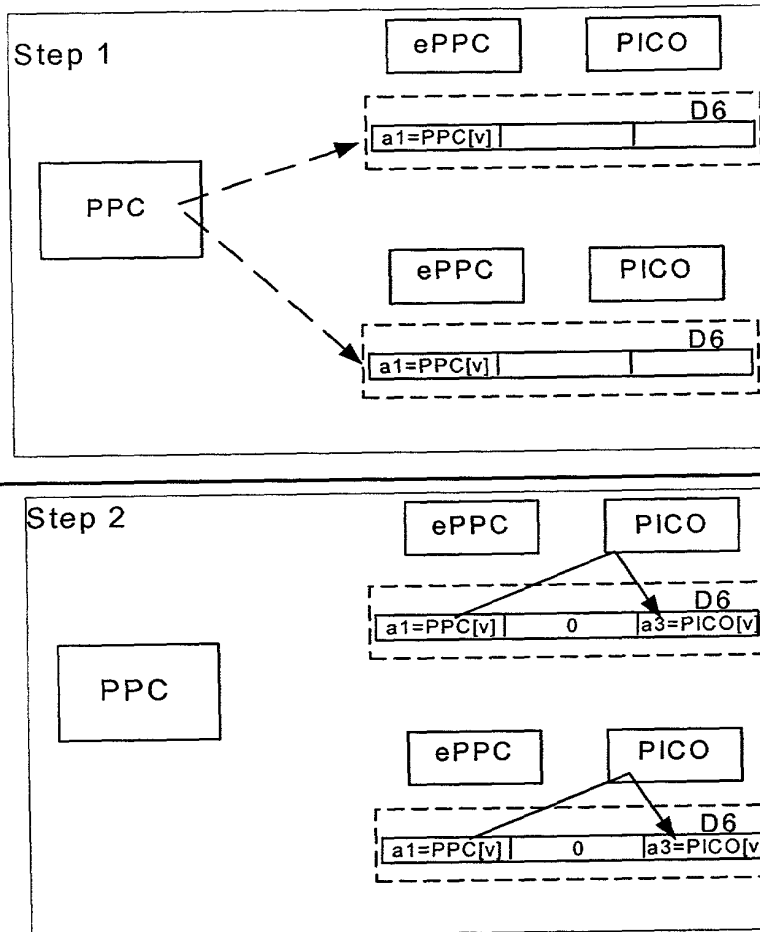
Step 3

$majority(a1,a2,a3) = majority(v,0,v) = v$, Receiver fault

Figure 20 - Receiver fault (doesn't relay a value)

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 20 of 36
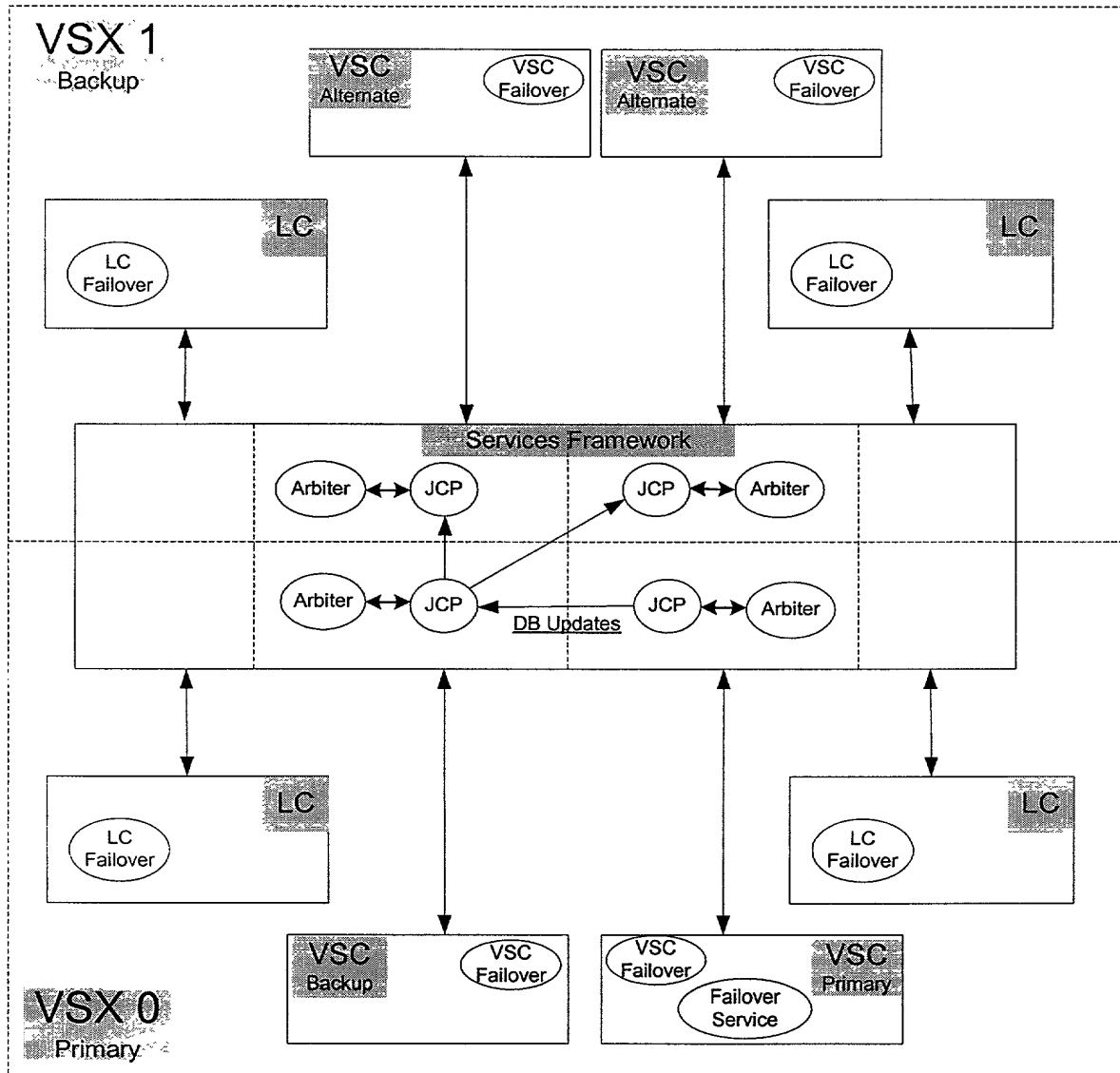
Figure 21A - Failover Service Architecture

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 21 of 36
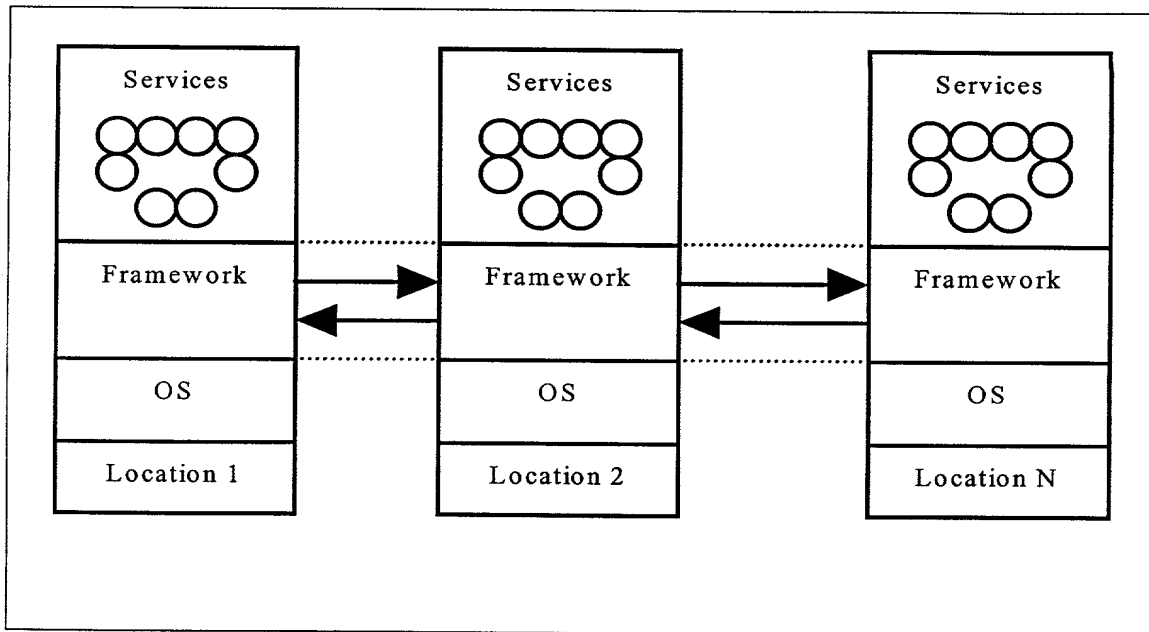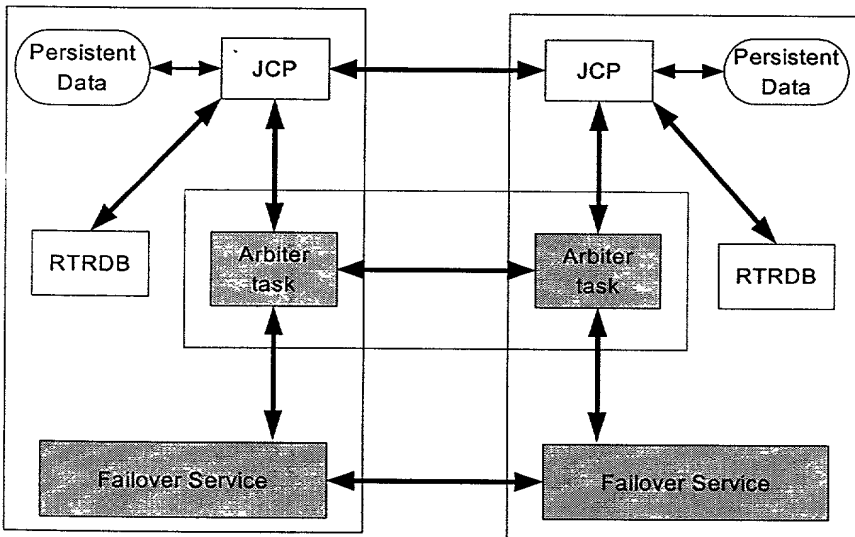
Figure 21B

Figure 22 - An Arbiter for the Database

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 22 of 36
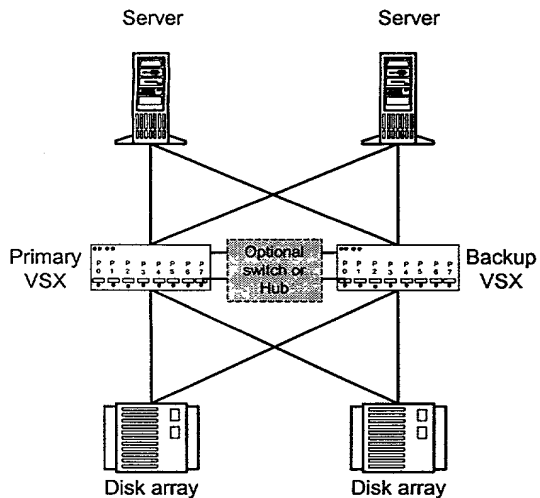
Figure 23 – Shared Link
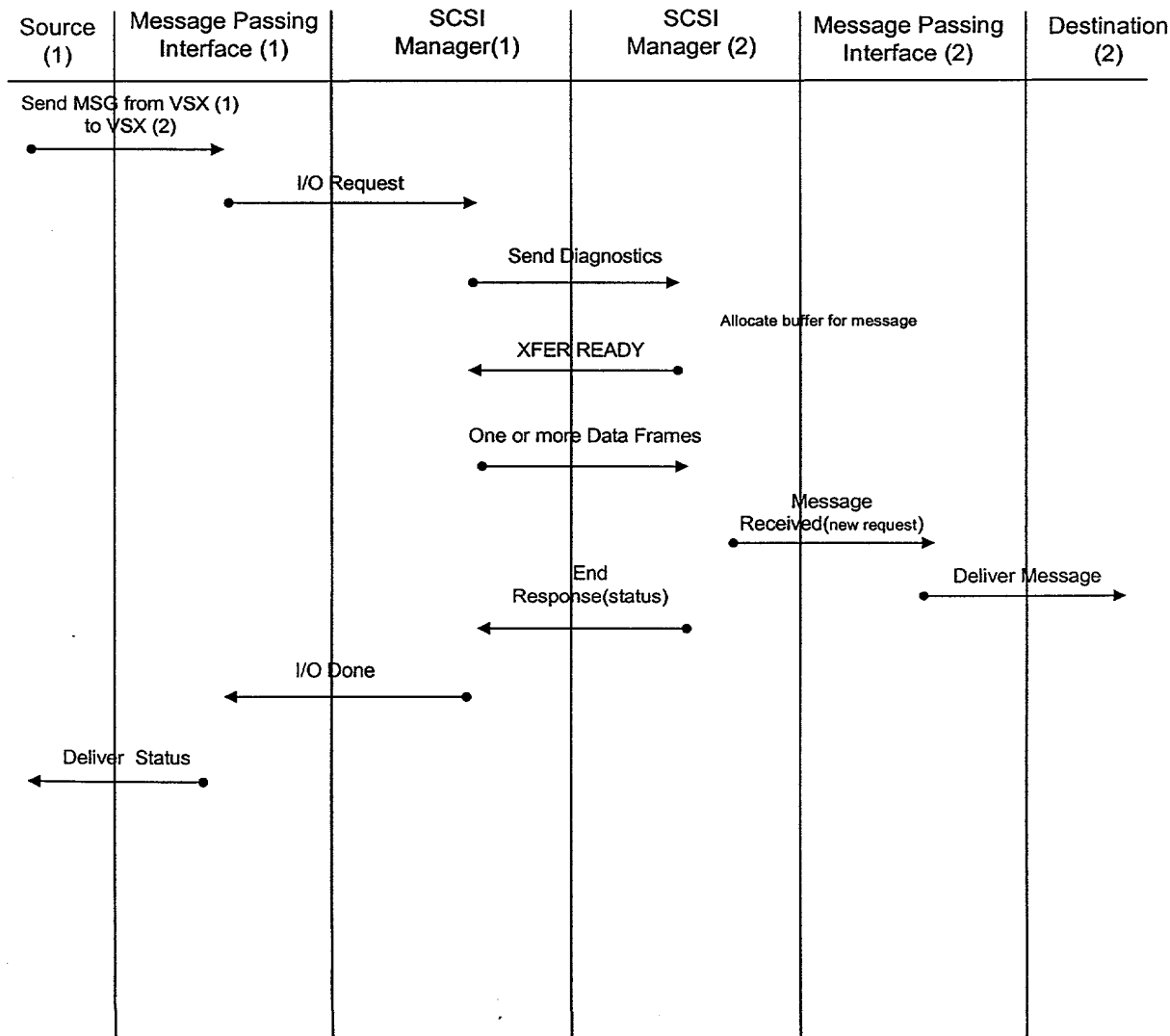
Figure 24 – VSX to VSX Message Passing

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 23 of 36
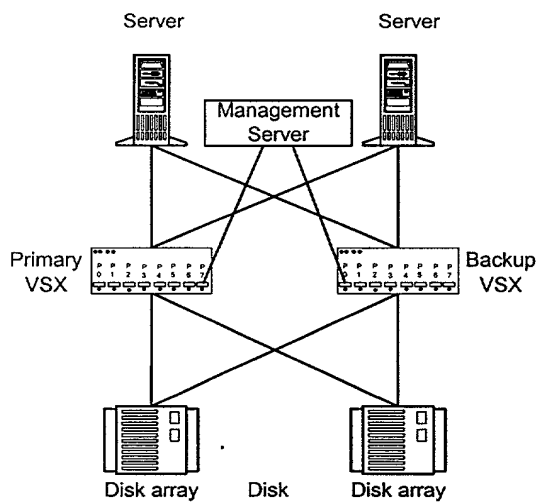
Figure 25 – Management Link
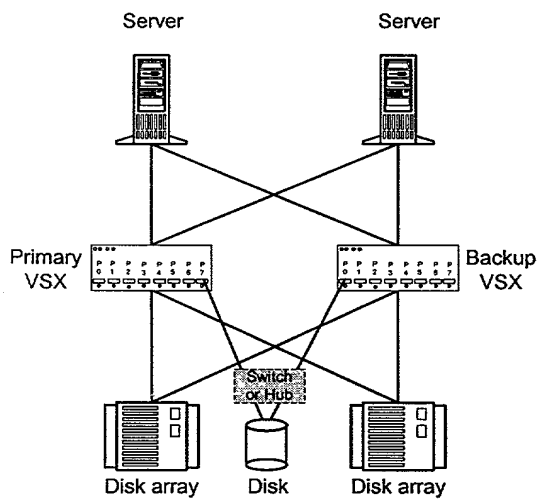
Figure 26 – Shared Disk

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 24 of 36

Source (1)	Message Passing Interface (1)	SCSI Manager(1)	SCSI Manager (2)	Message Passing Interface (2)	Destination (2)

Send MSG from VSX (1) to VSX (2)

I/O Request

Write Buffer

End Response(status)

I/O Done

Get Message [polls]

I/O Request

Read Buffer

End Response(status)

I/O Done

Deliver Message

Figure 27 – VSX to VSX communication Using Shared Disk

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 25 of 36
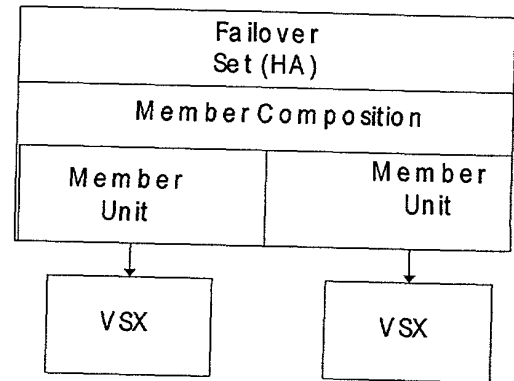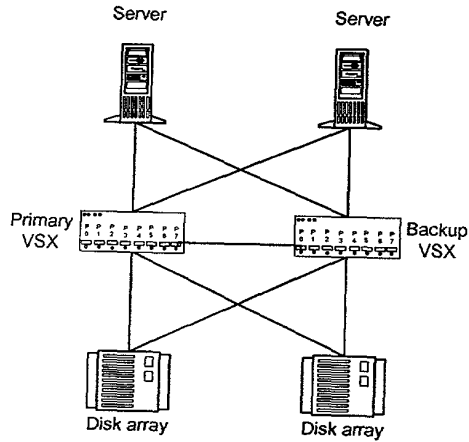
Figure 28 -2 Node HA Configuration
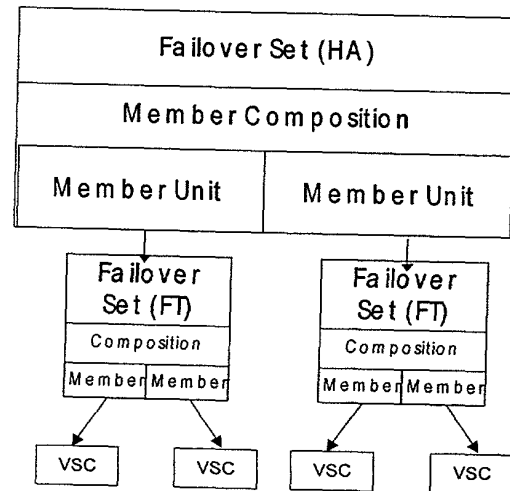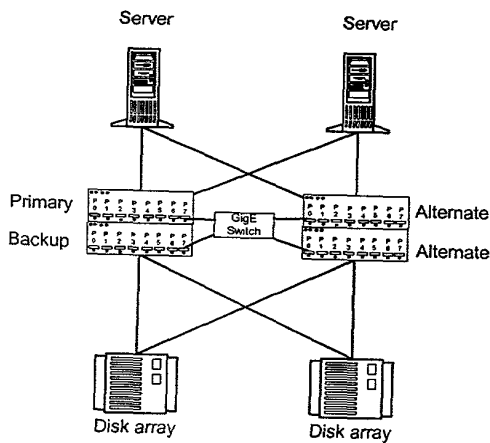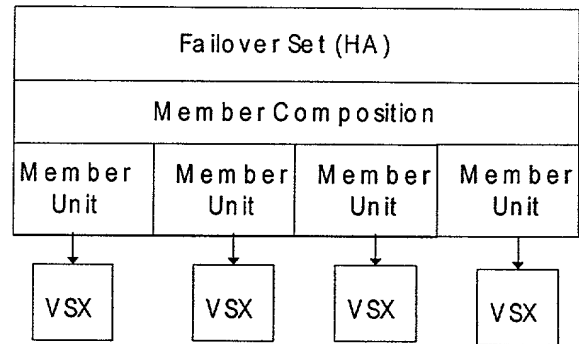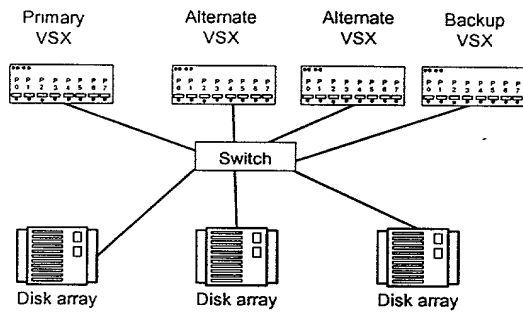
Figure 29 - Hierarchical HA Configuration

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 26 of 36

Figure 30 - N + 1 Nodes

Figure 31 - N – Nodes

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone. 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 27 of 36
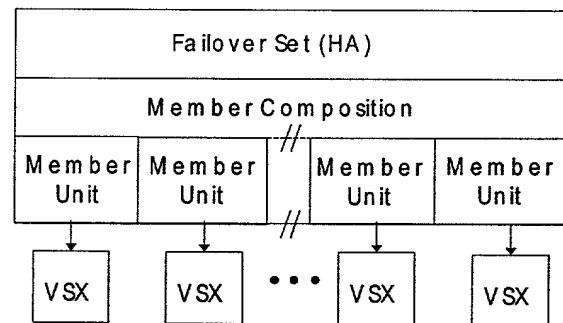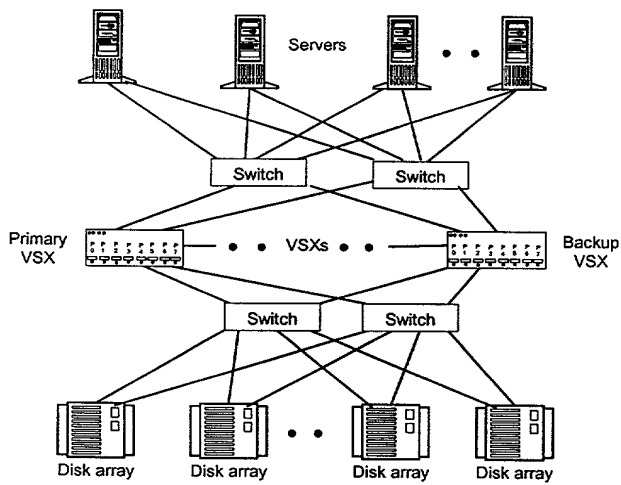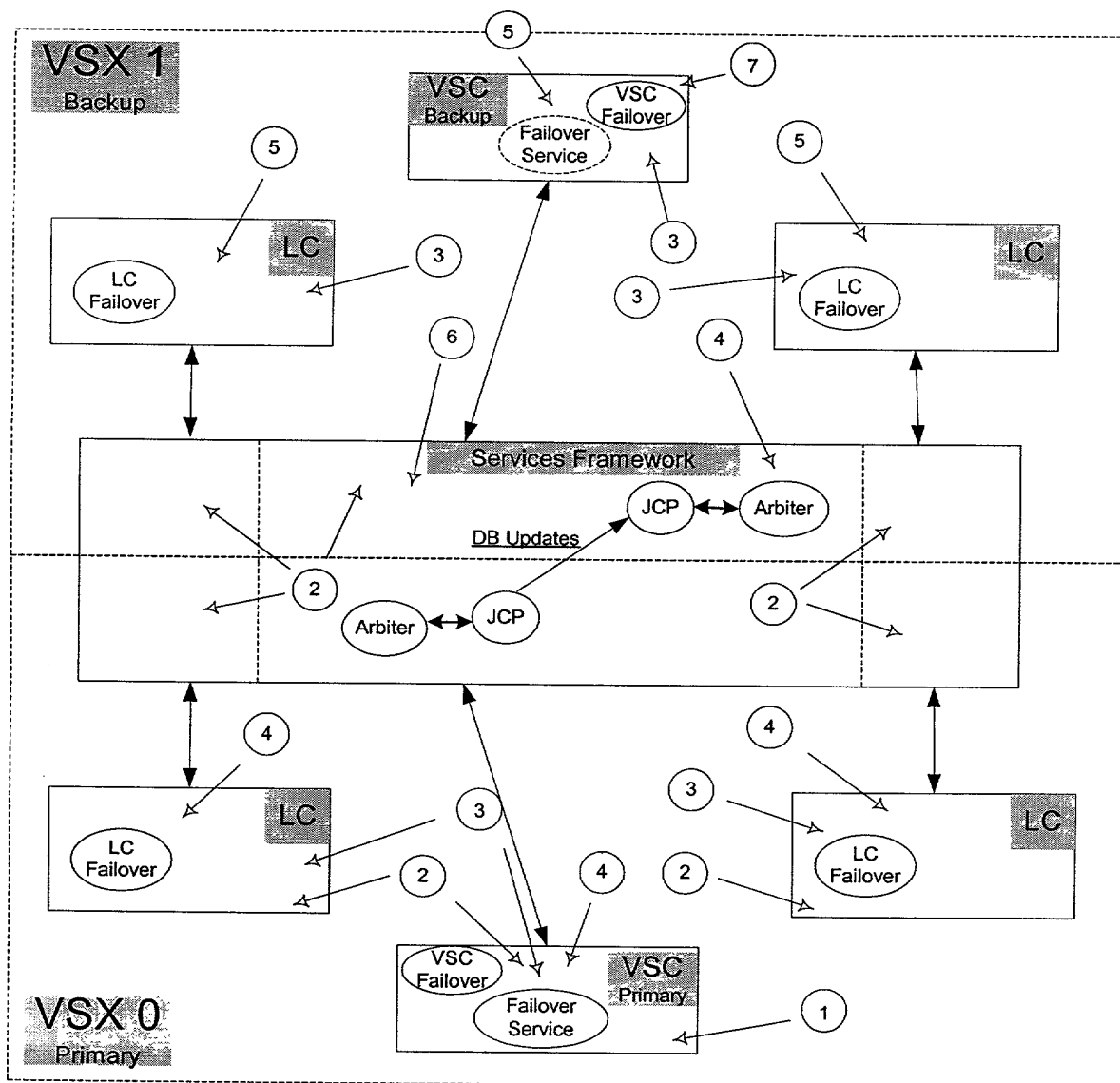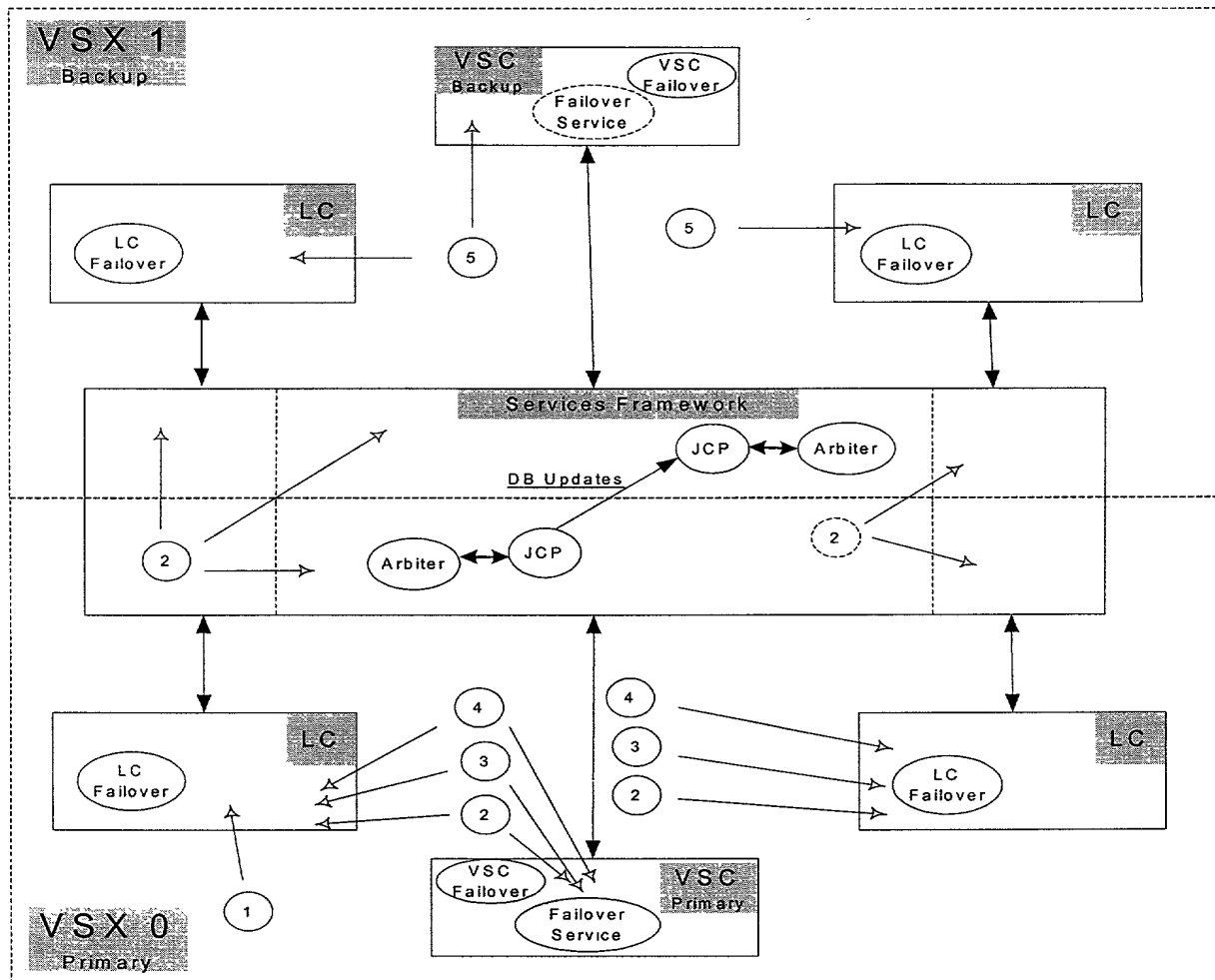
1. VSC Crashes (Host Processor)
2. Rest of system detects VSC crash
3. Error Analysis determines Member fails, which translates into a "Primary Lost" event
4. Activate JCP in Master mode and enable the virtual services, Stop Ports on failed Primary
5. Reset affected devices, Cleanup reservations and locks, Set Unit Attention
6. Restart management requests
7. Restart RCON and FORMAT

Figure 32 - VSX Failover, Primary Fails

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
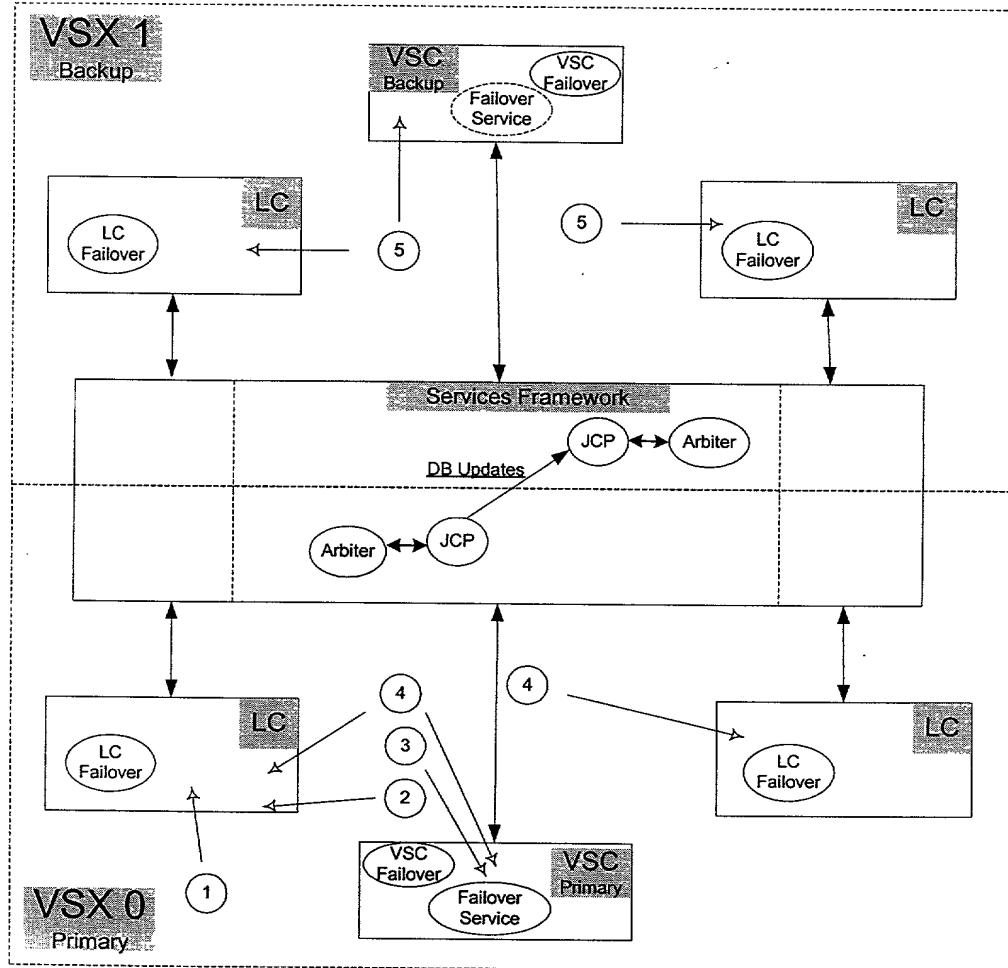Sheets of drawings 28 of 36

1. LC Crashes (Host Processor)
2. Rest of system detects LC crash
3. Error Analysis determines IO Path fails for all devices (server and storage) on LC
4. Upstream hLUNs report CHECK CONDITION for all devices connected to ports on failed LC.
   RCON and FORMAT aborted, if necessary.
5. Restart RCON and FORMAT, if necessary

Figure 33 - IO Path Failover - LC Fails

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
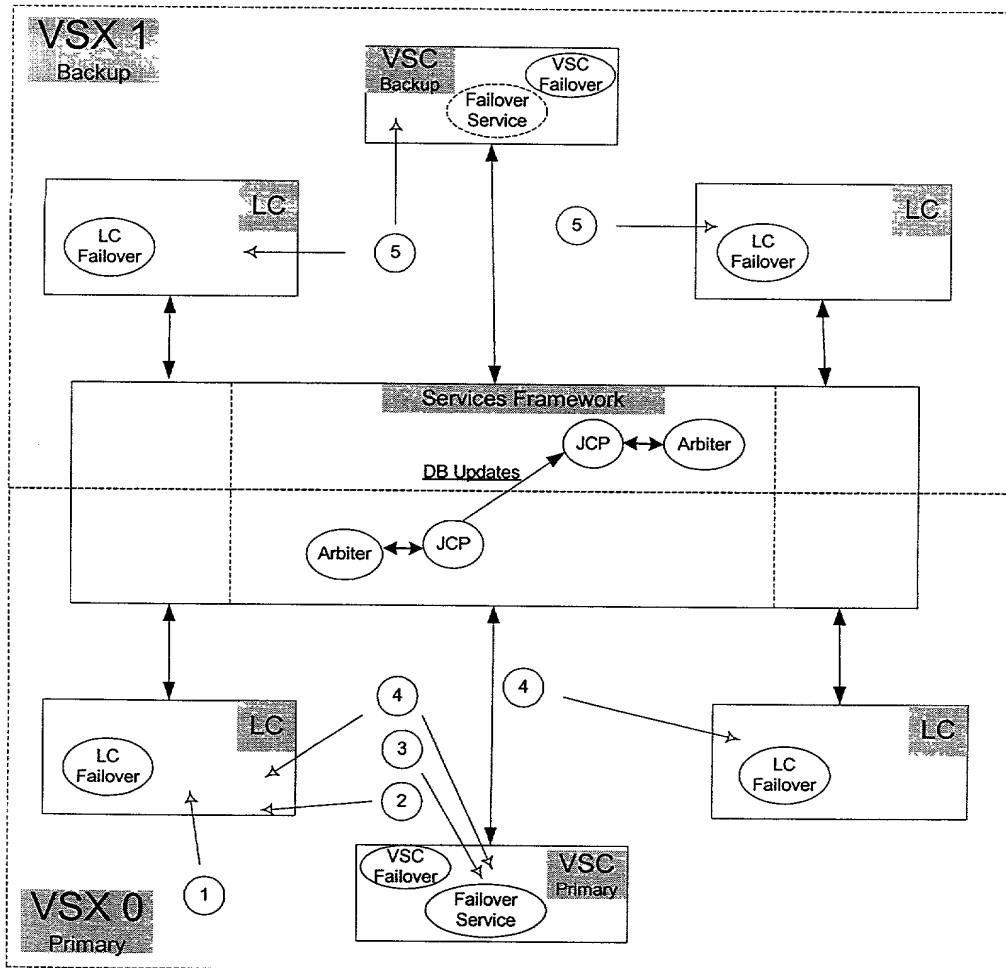Sheets of drawings 29 of 36

1. FC ASIC Crashes
2. LC detects FC ASIC crash
3. Error Analysis determines IO Path fails for all devices (server or storage) on FC ASIC
4. Upstream hLUNs report CHECK CONDITION for all devices connected to failed FC Ports. RCON
   and FORMAT aborted, if necessary
5. Restart RCON and FORMAT, if necessary

Figure 34 - IO Path Failover - FC Port Fails

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 30 of 36

1. Link down on port
2. LC detects FC Port link down
3. Error Analysis determines IO Path fails for all devices (server or storage) on FC Port
4. Upstream hLUNs report CHECK CONDITION for all devices connected to FC Port. RCON
   and FORMAT aborted, if necessary
5. Restart RCON and FORMAT, if necessary

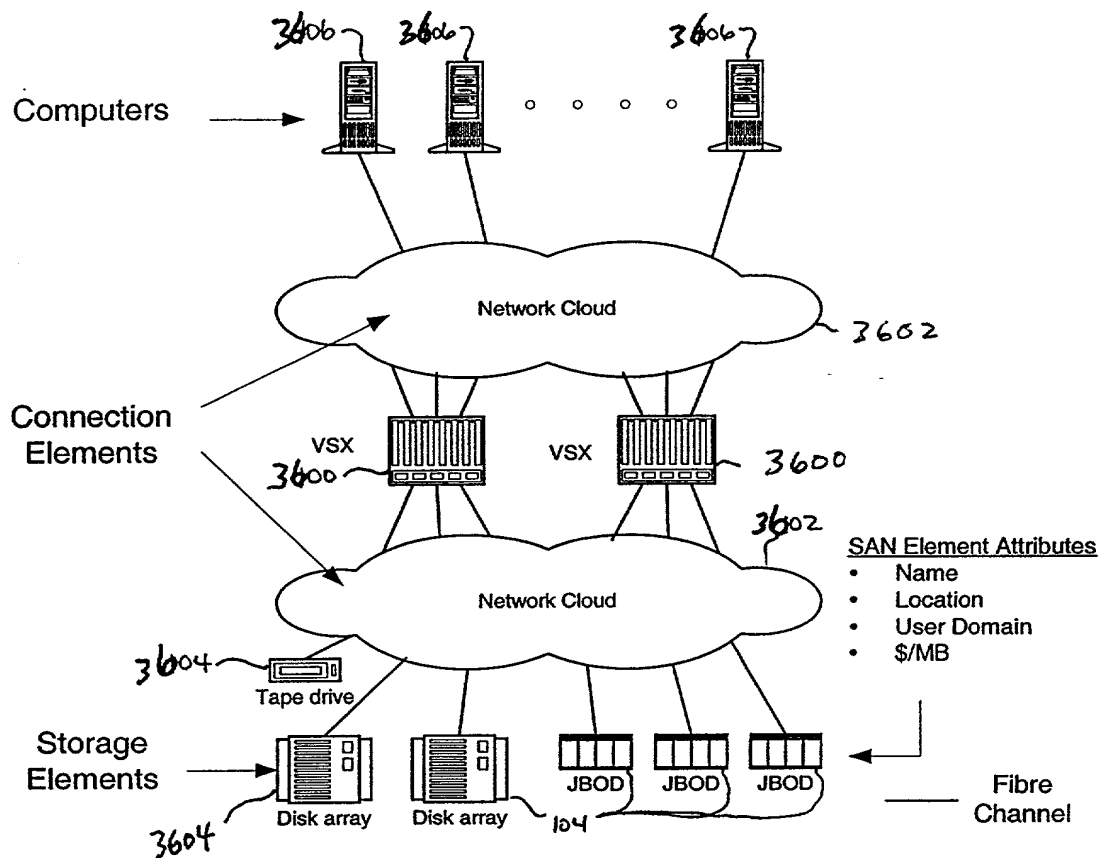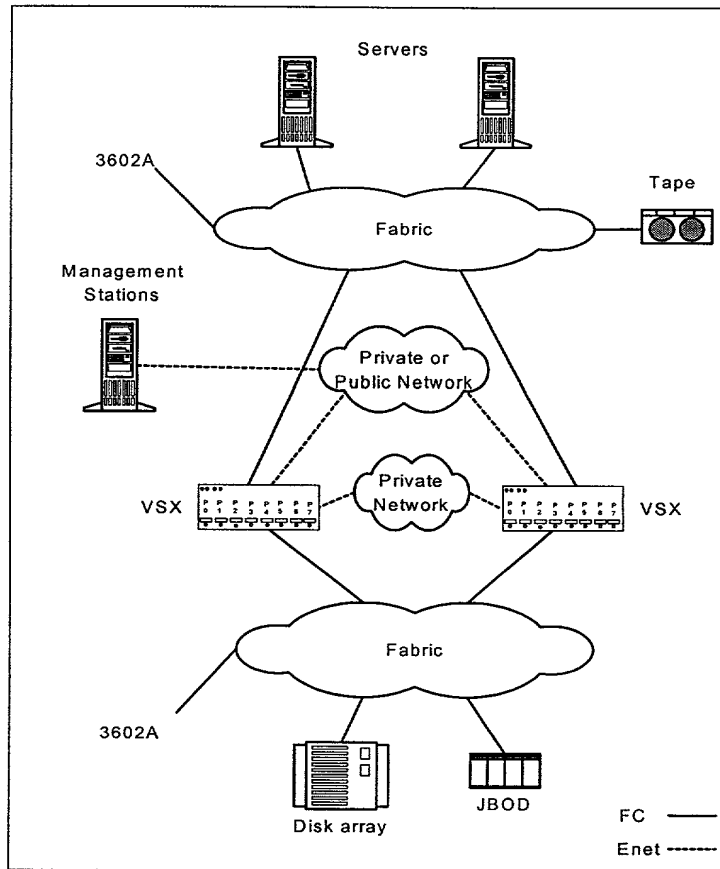**Figure 35 - IO Path Failover - Link Down**

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 31 of 36

Computers

3606 3606 3606

o o o o

Network Cloud 3602

Connection Elements

VSX 3600 VSX 3600

3602

SAN Element Attributes
- Name
- Location
- User Domain
- $/MB

Network Cloud

3604 Tape drive

Storage Elements

Disk array Disk array JBOD JBOD JBOD

3604 104

Fibre Channel

Fig. 36

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 32 of 36

Figure 36 A Physical Setup for VSX-HA – Variation 1

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 33 of 36
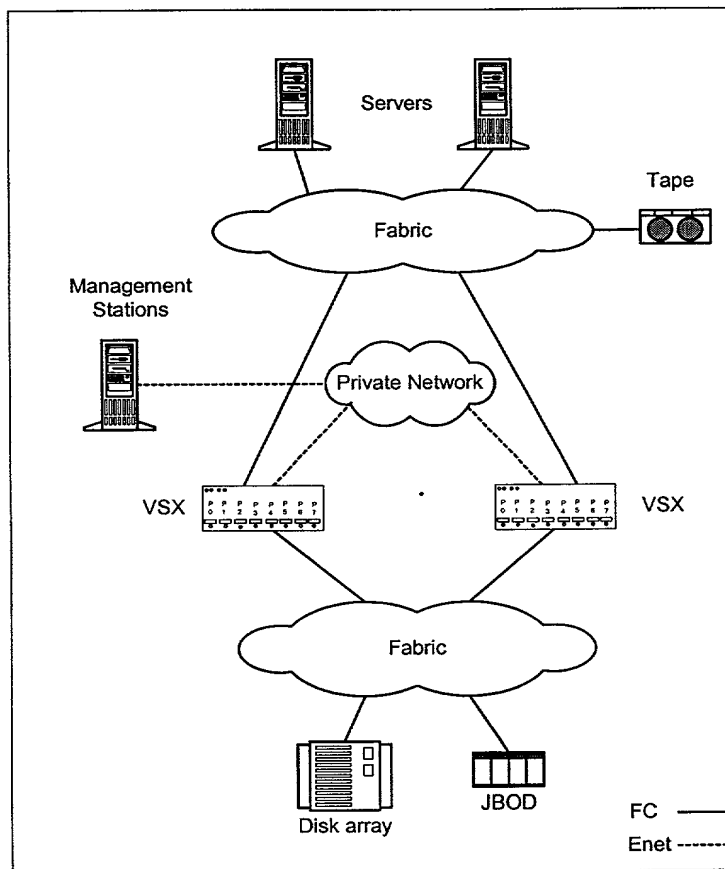
Figure 36 B Physical Setup for VSX-HA – Variation 2

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 34 of 36
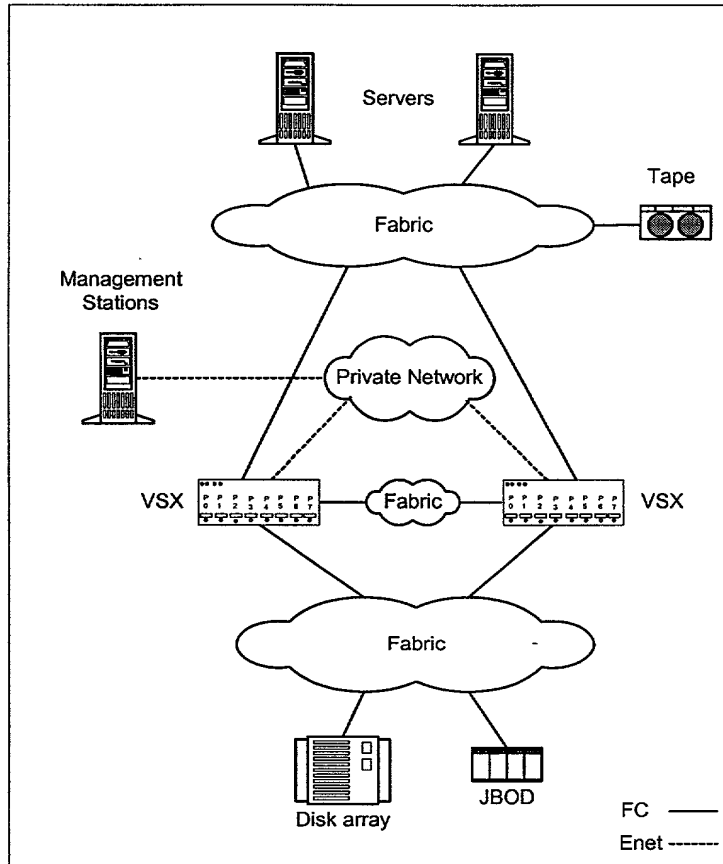
Figure 36 C Physical Setup for VSX-HA -- Variation 3

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 35 of 36
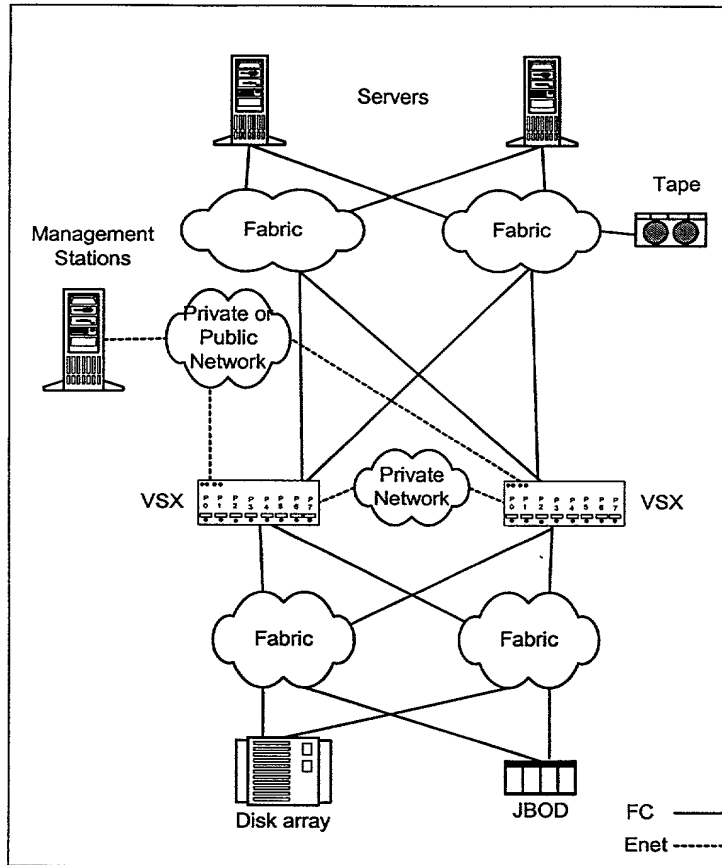
Figure 36 D Physical Setup for VSX-HA – Variation 4

20949P-000800US
George B. F. Yee, Reg. No. 37,478
Telephone: 650-326-2400
Inventors: Richard Meyer et al.
Title: Failover Processing in a Storage System
Sheets of drawings 36 of 36

Fig. 37